POSTGRADUATE SCHOOL
3940

NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

THE METHODOLOGY OF COST ESTIMATION
FOR U.S. MISSILES

By

Kyung Ho Choo

December 1982

Thesis Advisor: M.G. Sovereign

Approved for public release; distribution unlimited

T207840

LIBRARY, NAVAL POSTGRADUATE SCHOOL
MONTEREY, CA 93940

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): The Methodology of Cost Estimation for U.S. Missiles	5. TYPE OF REPORT & PERIOD COVERED: Master's Thesis; December 1982
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Kyung Ho Choo	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Naval Postgraduate School, Monterey, California 93940	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Naval Postgraduate School, Monterey, California 93940	12. REPORT DATE: December 1982
	13. NUMBER OF PAGES: 63
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

CER
Adjustment
Constant Dollar
Regression
Explanatory Variable
Prediction Interval
Standard Error
Learning Curve

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Public data on U.S. missile system are used to demonstrate the procedures and techniques for development of Cost Estimating Relationships(CER) by statistical methods. First, attention is given to data adjustment for constant dollars and quantities since the data come from yearly budgets. Next, simple and multiple linear regressions are performed in various combinations of the three explanatory variables (weight, speed and range).

20. (continued)

Learning curves are introduced to derive the reduction in cost as the number of items produced increases.

Approved for public release, distribution unlimited

The Methodology of Cost Estimation for U.S. Missiles

by

Kyung Ho Choo
Captain, Korean Army
B.A., Korean Military Academy, 1975

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN MANAGEMENT

from the

NAVAL POSTGRADUATE SCHOOL
December, 1982

ABSTRACT

Public data on U.S. missile systems are used to demonstrate the procedures and techniques for development of Cost Estimating Relationships (CER) by statistical methods. First, attention is given to data adjustment for constant dollars and quantities since the data come from yearly budgets. Next, simple and multiple linear regressions are performed in various combinations of three explanatory variables (weight, speed and range). Learning curves are introduced to derive the reduction in cost as the number of items produced increases.

TABLE OF CONTENTS

I. INTRODUCTION 9

II. DATA COLLECTION AND ADJUSTMENT 12

A. DATA COLLECTION 12

1. Data Collection 12

2. Observation of Data for Homogeneity . . . 15

3. Selection of Independent Variable(s) . . . 16

B. DATA ADJUSTMENT 17

1. Cost Definition Adjustments 17

2. Price Level Adjustments 17

3. Cost Quantity Adjustments 19

III. STATISTICAL METHOD FOR COST ESTIMATING RELATIONSHIPS 21

A. SIMPLE LINEAR REGRESSION 26

1. Least-squares Estimating 27

2. Statistical Inference 31

3. Prediction Intervals 34

B. MULTIPLE REGRESSION 38

IV. THE LEARNING CURVE 43

A. THE RELATIONSHIP BETWEEN COST AND QUANTITY . . 44

1. Log - linear Unit Curve 45

2. Log - linear Cumulative Average Curve . . 45

B. APPLICATIONS 46

C. EXAMPLE OF LEARNING CURVE FROM BUDGET DATA . . 47

V. CONCLUSION 50

APPENDIX A: EXAMPLE OF CALCULATIONS FOR THE LEARNING CURVE 53

APPENDIX B: MISSILE CLASSIFICATION 54

APPENDIX C: DESIGNATOR LISTING 56

APPENDIX D: LAUNCH ENVIRONMENT SYMBOLS 57

APPENDIX E: MISSILE SYMBOLS 58

APPENDIX F: VEHICLE TYPE SYMBOLS 60

LIST OF REFERENCES 62

INITIAL DISTRIBUTION LIST 63

LIST OF TABLES

I. Price Index 18

II. Twenty - two Missile Sets 23

III. Comparison of Multiple-linear with Simple-linear Regression Results 40

IV. Plot Point of RIM - 67B 48

LIST OF FIGURES

3.1 Scatter Diagram of Cost vs Weight for Data . . . 23

3.2 Scatter Diagram of Cost vs Speed for Data . . . 24

3.3 Scatter Diagram of Cost vs Range for Data . . . 24

3.4 Regression Line and Standard Error of Estimate . 29

3.5 The 90-percent Prediction Interval Band for Estimated Based on Sample Data 36

4.1 The Learning Curve of RIM-67B 49

I. INTRODUCTION

A cost estimate is a judgement or opinion regarding the future cost of an object, commodity, or service [Ref. 1: p.1]. In particular, in this thesis it is the cost of missiles. This judgement or opinion may be arrived at formally or informally by a variety of methods, all of which are based on the assumption that experience is a reliable guide to the future. In some cases the guidance is clear and unequivocal. In others, it is not. Much, perhaps most, estimating involves the relationship between past experience and future application. The more interesting problems are those in which the relationship is unclear, because the proposed item differs in some significant way from its predecessors. The challenge to cost analysts concerned with military hardware is to project from the known to unknown, for example, to use experience on existing missiles to predict the cost of the next-generation missile. The techniques used for estimating hardware cost range from intuition at one extreme to a very detailed "bottom-up" application of labor and material industrial engineering standards at the other. There are many methods to estimate costs, but this thesis will discuss only the statistical approach to estimating the cost of U.S. missiles.

In the statistical approach, estimating relationships that use explanatory variables such as weight, speed, range, and thrust are relied upon to predict the cost at a high level of aggregation, either the missile itself or major subsystems. To say that statistical techniques can be used in a variety of situations does not imply that the techniques are the same for all situations. They will vary according to the purpose of the study and the information available.

In a conceptual study, it is necessary to have a procedure for estimating the total expected cost of a program, and this must include an allowance for the contingencies and unforeseen changes that seem to be an inherent part of most development and production programs. In effect, this procedure merely asserts the obvious: as more is known, fewer assumptions are required. When enough is known, and this means when a product is well into production, accounting information and data can be taken directly from records of account and used with a minimum of statistical manipulation, i.e., only the adjustment for change in "learning" and inflation as the systems are produced. This technique is useful on those cases when the future product or activity under consideration is essentially the same as that for the past or current period, which is often not the case. But all new missiles vary in their characteristic parameters.

In any situation the estimating procedure to be used should be determined by the data available, the purpose of estimate, and, to an extent, by such other factors as the time available to make an estimate. In fact, since the life of a modern weapon system may run twenty years (or longer), the investment needed to establish a new system may be dwarfed by the costs required to operate and maintain it.

II. DATA COLLECTION AND ADJUSTMENT

The analysis of past cost data yields estimates of future costs based on the cost relationships of previous periods. The degree to which these data are appropriate depends upon the extent to which cost behavior in the future will correspond to that in the past and to the extent we identify the relationships. If the change being considered is extensive enough to bring about changes in the underlying cost structure such as the use of new technology, the unadjusted historical cost data may be inappropriate.

A. DATA COLLECTION

There are three steps in data collection.

1. Data Collection

"Data collection is the process of identifying, searching out, acquiring, verifying, and recording the specific information that is of value to the analyst." [Ref. 2: p.11] The cost analysts have many data sources. The cost information report (CIR) was established by DOD in 1966 to simplify the data collection problem. This reporting system was designed to collect cost and related data on major contracts for aircraft, missiles, and space programs.

Efforts are presently underway to enlarge the coverage of the CIR's to other areas of defense contracting. The new system is called contract cost data reporting, (CCDR). The reports are sent by contractors to the OSD's cost analysis improvement group (CAIG). In the absence of CIR-type data, the analyst must resort to contractor records, such as the cost performance reports (CPR), sent to government program managers, engineering records, managerial records, and other periodical reports containing cost data such as the CSFR (Contractors Status of Fund Report). But for sub-systems, this type of data is not necessarily available to the analyst.

While collecting data, the analyst should keep in mind the levels of accuracy and aggregation that he needs. If cost data is available down to the component level, it may be possible to proceed with a disaggregated method of cost estimating, estimating each component and then aggregating. The advantage is no matter what approach is used, data collection problems can be minimized by first becoming familiar with the system's technology and second, by using consistent definitions for the cost and parameteric variables. For example there are at least three different types of historical data required to develop a statistical cost-estimating procedure. First, there are the resource data, usually in the form of expenditures and labor hours. It is customary to apply the word cost to both, and that

practice is followed throughout this thesis. A second type of data describes the possible cost-explanatory elements; for hardware such as aircraft and missile this means performance and physical characteristics. The third type is program data, i.e., information related to the development and production history of past hardware programs.

a. Resource Data

Resource data are generally classified under end-item categories or functional categories. An example of the former in various possible levels of detail are system, subsystem, component, and part. The functional cost categories, such as engineering, tooling, manufacturing, quality control, purchased equipment, are usually broken down into cost- elements--labor, material, overhead, and other direct charges. The data source is the contractor's plant. Generally, the accounting systems will vary from one company to another and the amount of detail is immense. Theoretical considerations aside, estimating techniques must be based on whatever resouce data the analyst can find, and in the past the availability of data has varied from one kind of equipment to another. The most data is given in the CPR which is currently available only by going to each project's management.

b. Physical and performance characteristics

Information about the physical and performance characteristics of missile system is just as important as resource data. Data collection in this area can be time-consuming, particularly since it is not often clear in advance what data will be required. The goal, of course, is to obtain a list of those characteristics that best explain difference in cost. Weight is a commonly used explanatory variable, but weight alone is seldom enough; speed is almost always included as a second explanatory variable for missiles or aircraft. But speed is often useful only at the total system level.

c. Program Data

A third type of essential data is drawn from the development and production history of hardware items. The acceptance data of the item, the significant milestones in the development program, the production rates, and the occurence of major and minor modifications in production--all such information can contribute to the development of cost-estimating relationships. The schedule data are needed for price adjustment in this thesis, for example.

2. Observation of Data for Homogeneity

Data must be checked to ensure that the cost changes reflect only changes in the selected explanatory variables.

If changes have occured from period to period in technology, skills of the labor force, or the price level of inputs, the cost measurement will be an amalgam of the change in output and the changes in design characteristics in the environment. Thus the cost data will not be homogeneous from observation to observation. Nonhomogeneous observations often result from technological or organizational differences in different plant or producing nearly identical output. In order to work with a large number of observations covering a wide range of output, it may be necessary to work with the cost data of many similar departments. If the nature of operations of the department varies, the behavior of the costs will reflect this diversity. Unadjusted cost data should not be used if these differences are significant. One solution is to aggregate the data above organizational differences. Another is to add an explanatory variable that measures the difference.

3. Selection of Independent Variable(s)

This step is analogous to the model building stage in any research project. While the cost relationship will usually be simple, involving only a few independent variables, it must be hypothesized before the analysis can be carried any further. Generally, we should choose an independent variable on the basis of a reasonable belief that some relationship exists between the variable and the cost

being estimated. The variables used in the estimating should be the ones that exert the major effect on the cost observed. Among the most widely employed variables are weight, and speed.

B. DATA ADJUSTMENT

There are three kinds of data adjustments.

1. Cost Definition Adjustments

Different contractor accounting practices and types of contracts are the primary reasons for this type of adjustment. An analyst should state the cost definition that he wishes to use and then adjust the data to meet his definition. It is sometimes impossible to obtain information needed for consistent adjustments. Interpretation of the final cost estimate should make allowances for this possible source of anomalous cost behavior.

2. Price Level Adjustments

It is all too apparent that inflation changes the purchasing power of the dollar dramatically. In order to compare the cost of a system purchased in 1953 to the cost of a new system, the cost figures must be adjusted to "constant" dollars. The Bureau of Labor Statistics publishes many indices that can be used for this purpose. With sufficient data, it is possible to produce a weighted index

TABLE I

Price Index

Price Index Base - 1983	
Year	Index
1972	39.79
1973	41.42
1974	45.09
1975	51.17
1976	54.95
1977	58.73
1978	62.95
1979	68.80
1980	78.79
1981	86.23
1982	93.20
1983	100.00
1984	106.21

Source: OSD (COMPTROLLER) 1982

specifically for the type of system being estimated. This can be a very laborious process and so several general indices are available for use. The producers price index (PPI) is most useful for constructing indices for the various appropriation accounts used by the military

services [Ref. 3: p. 24]. The Department of Defense also publishes a procurement index to be used for general military hardware. Best results are obtained from indexes which are specialized to the type of equipment being estimated. It is almost an impossibilty to obtain an index that will remove all of the price level changes of a particular item. Table I gives the index needed to adjust the missile costs in this thesis.

3. Cost Quantity Adjustments

The "learning curve" is a phenomenon prevalent in many industries. As the cumulative number of identical items produced doubles, the unit cost or a cumulative average cost is reduced by a constant percentage showing "learning".

Learning curve information can be obtained from two possible sources. The best source is the contractors cost records or CIR-reports for individual units. Costs of the units are plotted and a line is fitted to the plotted data. A second source of information would be a general industry-wide learning rate that may be published in the industry's literature.

If a general learning rate is available, say 90 %, along with the cost of a particular unit (say unit #5), the curve can be drawn by computing the cost for unit #10 (unit #5 cost times .9), plotting the two points, and drawing a line connecting the points on log-log paper. The assumed

learning curve can contribute large dollar errors if the assumed rate used is not accurate. For example a +1 % learning rate error in the example above gives a 3.261 % difference over 40 units. Quantity adjustment will be discussed for the missile data in Chapter IV.

III. STATISTICAL METHOD FOR COST ESTIMATING RELATIONSHIPS

Cost estimation relationships (CER) are developed from the historical cost of like systems and the parameters (e.g., weight, maximum speed, range.) of these systems.

Statistical amalysis can help provide an understanding of factors that influence cost, but estimating relationships are not a substutite for understanding: regression analysis, which will be discussed in this thesis, does not offer a quick and easy solution to all the problems of estimating cost. The outstanding characteristic of a CER is that the relationship between cost and explanatory variable is direct and obvious; thus, cost per kg (or pound) is widely used because of the generally satisfying thesis that as missile, tank, or airplane increases in weight it becomes more costly. Weight changes alone do not always adequately explain cost changes, and additional explanatory variables are often needed. The problem is to find these variables and their relationship to cost. The producer is to decide what variables are logically or theoretically related to cost and then to look for the patterns in the data that suggest a relationship between cost and the variables. Table II contains a set of data on cost and selected variables that can be analyzed for such patterns. The costs of twenty two missiles sets are given with weight, speed, and range of

TABLE II

Twenty - two Missile Sets

MISSILES	WEIGHT (kg) Launch	WEIGHT (kg) Pay load	SPEED (mach)	RANGE N mile	CUM AVE COST 1983 ($ Mill) (1,000th)
AGM - 86B	1,426.1	108.9	0.7	1350.0	1.9330
MIM - 72C	83.9	12.7	2.5	2.6	0.1039
FGM - 77A	11.5	2.5	0.3	0.5	0.0503
BGM - 109C	1,224.7	-	0.7	2000.0	3.1826
AGM - 88A	353.8	66.2	3.5	10.0	0.8632
RGM - 84A	630.0	231.3	0.8	60.0	1.1668
AGM - 114A	44.7	9.1	1.0	3.8	0.1496
MIM - 23B	623.7	30.7	2.5	25.0	0.7952
MGM - 52C	1,285.5	210.9	3.0	75.0	1.0599
AGM _ 65A/B	215.4	58.9	1.0	65.0	0.1554
AGM - 65D	215.4	58.9	1.0	65.0	0.3596
MIM - 104	902.7	90.7	3.0	37.0	3.1805
PERSHING-II	4,600.9	294.8	8.0	1000.0	3.3653
AIM - 54A	446.8	59.9	5.0	76.0	1.2326
AIM - 54C	453.6	59.9	5.0	100.0	1.2512
MIM - 115	63.5	5.9	1.6	5.0	1.1975
AIM - 9L/M	86.2	11.3	2.5	1.9	0.1329
AIM - 7F/M	231.3	40.8	2.5	24.0	0.5271
RIM - 67B	1,358.9	61.2	3.0	69.0	0.9878
RIM - 66C	640.0	61.2	3.0	40.0	0.6146
FIM - 92A	15.7	0.9	2.0	3.0	0.1275
BGM - 109A/B	1,224.7	453.6	0.7	300.0	2.5616

each. It is to be expacted that cost would increase with weight or with speed or range.

A graphic analysis of the data in Table II shows that cost is not a simple linear function of any of the three expalnatory variables. Cost tends to increase with weight, but there are notable exceptions to the trends, as illustrated by the scatter diagram of Fig. 3.1. Cost is plotted against speed and range as shown in Fig. 3.2 and Fig. 3.3. At this point, it is not clear if any of the explanatory variables, either singly or in combination, will yield a useful estimating relationship.

To illustrate techniques that are commonly employed in deriving estimating relationships, assume that cost can be

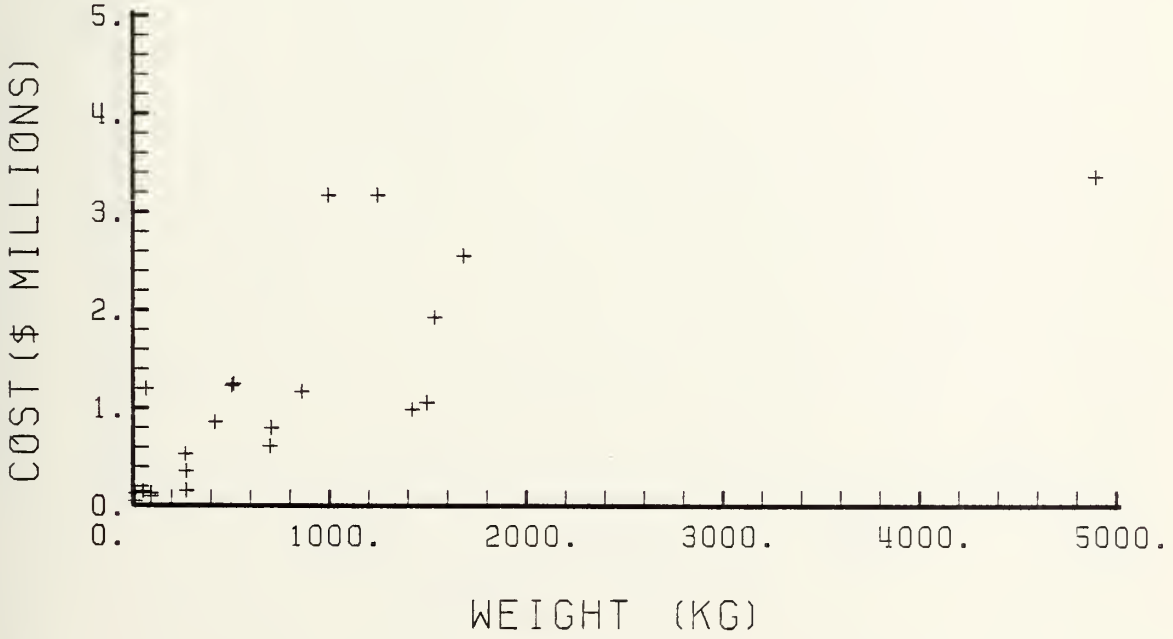


Figure 3.1 Scatter Diagram of Cost vs Weight for Data

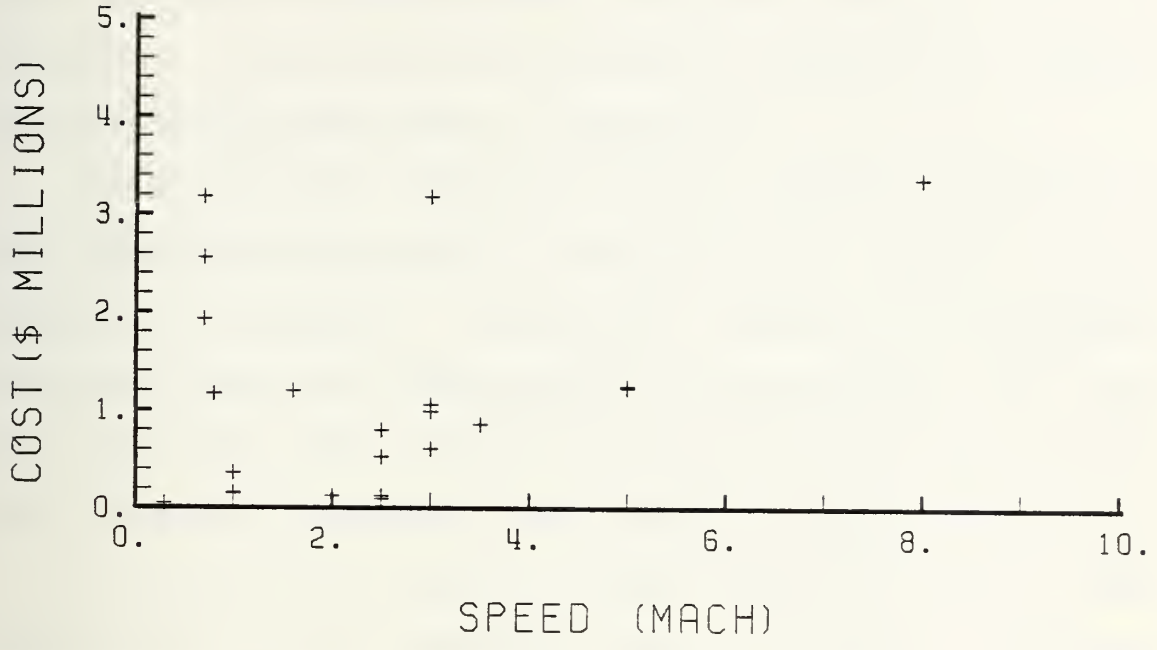


Figure 3.2 Scatter Diagram of Cost vs Speed for Data

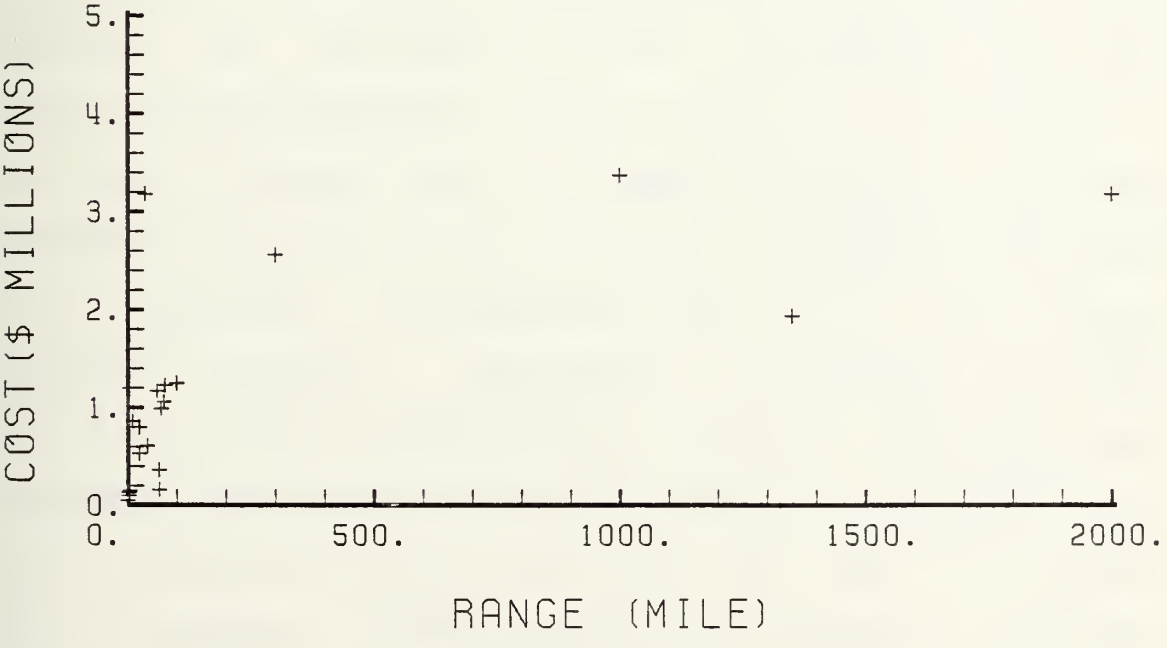


Figure 3.3 Scatter Diagram of Cost vs Range for Data

related to a single predictive variable--that of weight. The results of a simple linear regression model will then be examined. Later, several explanatory variables in multiple regression analysis will be considered.

The statistical technique normally applied to developing CERs from historical cost and parameteric data is called regression analysis. Regression analysis is primarily concerned with the determination of the equation of a line or curve which will predict how one variable (e.g., cost) will vary with respect to some parameter (e.g., weight).

Regression has become a widely accepted tool for cost analysis and it is frequently used to develop estimating relationship. The technique of regression analysis can be thought of as consisting of two distinct stages. The first is that of estimating the constant and coefficients of the equation, and the second is that of inferring the reliability and significance of the results of the estimate on the basis of assumed (and to a degree verifiable) properties possessed by the data and the results. Regression analysis as a technique is applicable only to the two stages performed together. Estimating coefficients or curve fitting is simply a mathematical exercise. Only when these estimating procedures are used as a basis for making statistical inferences can they be viewed as part of a statistical analysis. Before performing regression analysis, guidelines had to be established for determining what const-

itutes satisfactory regression criteria. The guidelines established for this missile study were as follows:

1. The interaction between the dependent cost variable and the independent variables (time-index value, etc.) is such that changes in the latter will generate reasonable changes in the former.
2. The number of variables will be limited to three or less because of the limited sample size.
3. Good statistical parameters such as low (20 percent or less) coefficient of variation, significant coefficient t-test, small standard error of estimate, etc. are achieved [Ref. 4: p. 10].
4. The relationships should be as applicable as possible to missile systems beyond the performance range of the sample data; i.e., future systems.

A. SIMPLE LINEAR REGRESSION

Scatter diagrams of the 1st "theoretical" unit versus the 1000th unit cost versus the time-index value of the 1000th unit were plotted and analyzed. But most analysts usually choose the 1000th unit as a better projection quantity than the 1st unit; the 1000th unit is the standard unit used in this thesis. The form of the relationship between cost and the explanatory variable(s) depends upon the problem. It may reflect an underlying physical form that is suspected. For physical characteristics, a simple linear

model is frequently used to describe the relationship between two variables. In this case, the equation of the model is

$$y = a + b\,x,$$

where y is the dependent variable and x is the explanatory variable. The system a and b are the constant and coefficient, respectively, of the equation estimated from the data. Here y could represent the procurement cost of missiles and x could represent the weight. If it is assumed that b is greater than zero, the model indicates that heavier equipment will cost more than lighter equipment. When the values of a and b are known, it is possible to estimate (cost) for any given value of x (weight).

1. Least-squares Estimating

Given equation $y = a + b\,x$, the basic problem in the first phase of the regression analysis is to derive estimates of the parameters a and b. The standard procedure is the method of least-squares. The values of a and b are determined by the requirement that the sum of the squared deviations of the sample observations from the estimated line will be minimum. Symbolically, This minimum is expressed as

$$\min \sum_{i=1}^{n} (y_i - \dot{y}_i)^2,$$

where y_i is the ith observation and $\dot{y}_i$ is the value of y_i

estimated from the equation

$$\dot{y}_i = \dot{a} + \dot{b} x_i$$

The dots over $\dot{a}$ and $\dot{b}$ indicate that $\dot{a}$ and $\dot{b}$ are least-squares estimates of the true but unknown values of a and b. Thus $\dot{y}_i$ is the least-squares estimate of y_i and the term $(y_i - \dot{y}_i)$ indicates the difference between each observed y_i and the corresponding estimated value $\dot{y}_i$. Figure 3.3 below contains the outcome of a least-squares regression performed on the data in Table II. The equation of the illustrated regression line is :

$$\dot{y} = 0.5353 + 0.7289 W$$

An analyst who obtained such a model should concerned with the question: How well does the equation fit the data?

There are several statistical measures that can give indication of the ability of the model to describe the data. The most commonly used measure of the "goodness of fit" of the regression equation is the coefficient of determination (r^2).

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}} = 51.3\ \%$$

The coefficient of determination is the percentage of the variation in the data explained by the regression model.

Ideally an analyst would want r to approach 1.00. The remaining variation may be explained when other

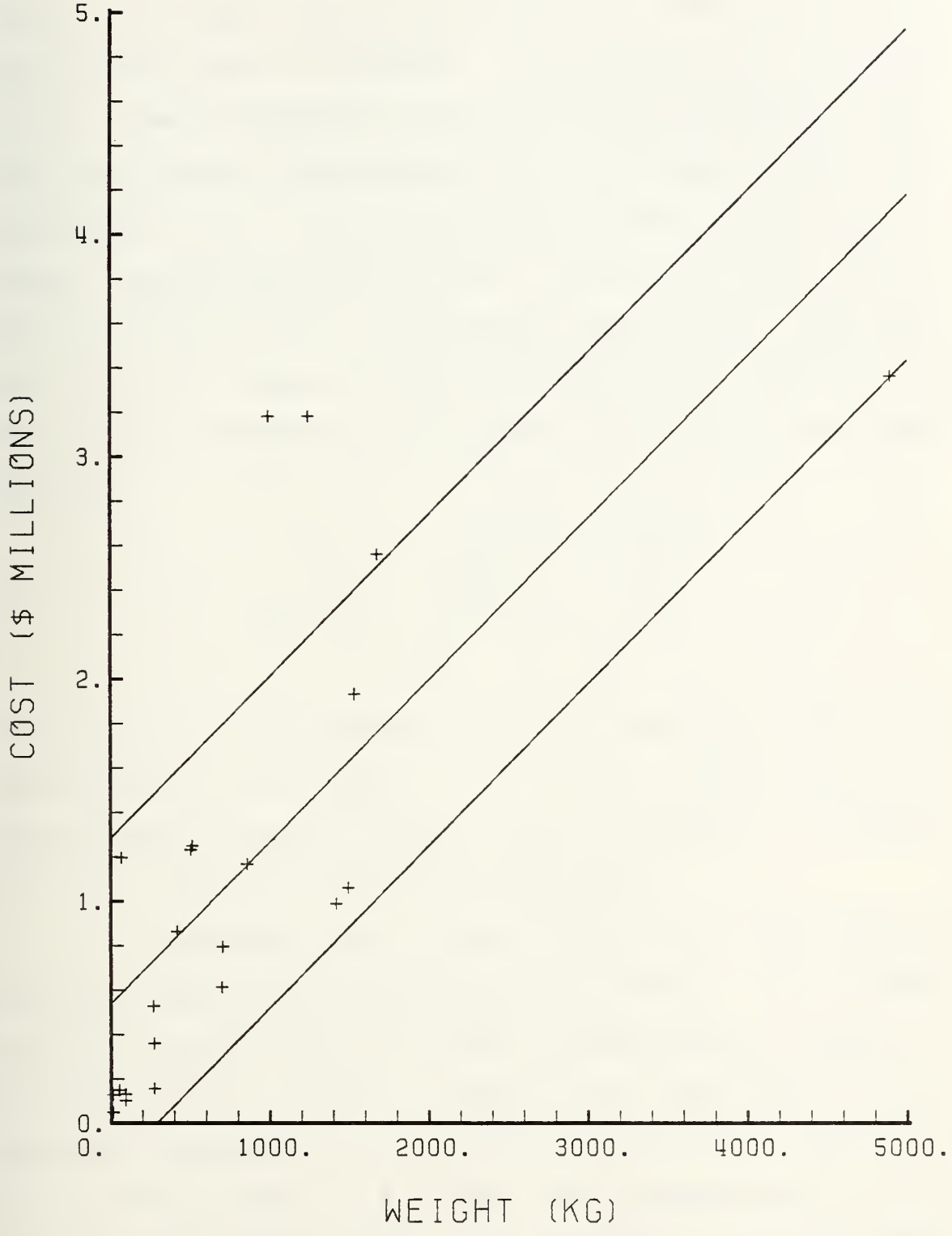


Figure 3.4 Regression Line and Standard Error of Estimate

variables are considered and brought into the equation. Figure. 3.4 also has plotted on it the lines representing the standard error of estimate. The greater the dispersion of the observed values of cost about the regression line, the less accurate the estimates that are based on that line are likely to be. If the cost data follows a normal distribution, approximately 68% of the data points should fall in the area bounded by the two standard error lines. The standard error of regression is a measure of the dispersion of the data and defined as the square root of the unexplained variance:

$$SE = \sqrt{\frac{\Sigma (Y_i - \dot{Y}_i)^2}{n - 2}}$$

This value of SE has been plotted above and below the regression line in Figure. 3.4. The interpretation and significance of these results will be discussed in connection with the use of prediction intervals.

In comparing one SE with another, it is useful to compute a relative standard error of estimate. The coefficient of variance (CV) is such a measure which relates the standard error of the model to the mean value of the dependent variable. A value of 10 to 20 percent for the CV is desireable [Ref. 1: p. 44]. The standard error of the model presented above is $ 0.7547 millions, and the coefficient of variance is: [Ref. 1: p. 44]

$$CV = \frac{S.E}{\bar{y}} = 0.6563$$

$\bar{y}$ = mean value of the dependent variable.

This 0.6563 value of the CV also serves as indication that the proposed model is not well suited to the data.

2. Statistical Inference

Statistical inference may be used to answer the two questions that arise in connection with the problem of reliability. To decide whether x and y are related, test for statistical significance; to evaluate predictions, establish a prediction interval for the regression line. However, certain assumptions and conditions must be met before standard techniques of statistical inference and testing can be validly applied to least-squares results; namely, the data are assumed to be a sample taken from a larger population, which meet the following conditions:

1. The x values are nonrandom (fixed) variables.
2. The residual deviations are independent random variables with normal distributions.
3. The expected value of the distribution of each of these random variables is zero, and the unknown variance is the same for all values of x.

Under these assumptions, the hypothesized relationship between y and x becomes:

$$Y_i = a + b x_i + u_i,$$

where $i = (1, \ldots, n)$,

u_i = the normally distributed random error term with zero expected value and a common and unknown variance.

Further, under these assumptions, the least-squares method produces unbiased maximum likelihood estimators. Standard statistical techniques can be applied to the least-squares results to test for significance and to make inferences about reliability and accuracy in a probabilistic sense. Although the subject of statistical testing is too complex to treat comprehensively here, the method of testing the significance of the relationship between x and y in the simple regression of Figure 3.4 will be examined briefly. Basically, the procedure involves establishing the null hypothesis that x and y are not related (i.e., that b=0), and testing to determine whether the hypothesis should be rejected. The test that is commonly used for this purpose is known as the t-test because it uses the t-ratio, or ratio of a coefficient to its standard error. For this simple regression, the ratio is expressed as

$$t_b = \frac{\dot{b}}{s_b} = 5.13$$

where $\dot{b}$ = the estimated regression coefficient (from the quation $\dot{y} = \dot{a} + \dot{b}x$),

s_b = The standard error of b,

$$= \frac{S\,E}{\sqrt{\Sigma(x_i - \bar{x})^2}}$$

S E = The standard error of regression.

A standard table of t-ratios is required to use t-ratio equation, to test the null hypothesis. If the calculated value of t_b falls below the appropriate value of t selected from this table, the null hypothesis that b = 0 would be accepted, and it would be concluded that b is, in fact, not significantly different from zero. The level of significance indicates the probability that the null hypothesis will be rejected when it is true. If there were evidence to justify the assumption that the sign of the coefficient could be only positive (or only negative) if it were different from zero, the level of significance associated with each t could be read directly from Student's t Critical Points Table. However, the common practice in regression analysis is not to make this assumption, but to test as though the value of t (if it were different from zero) could be either positive or negative. because of the distribution of the t-ratios, the level of significance for the two-sided test is twice the level of significance for the one-sided test. Thus, the

levels of significance of the t-values shows in the Student's t Critical Points Table are only half the actual levels for the two-sided test.

The question at this point is, what should the level of significance be for rejecting the hypothesis ? Unfortunately, no simple answer is possible. The values of .10, .05, and .01 are those that are most commonly used, but the analyst must make a decision based on the risk that is assumed when a true hypothesis is rejected. For this missile data no reasonable level would fail to reject the hypothesis that b = 0.

3. Prediction Intervals

The procedure for calculation of the prediction interval for a simple regression is as follows. For a given value of the explanatory variable, say x, the estimating equation is used to obtain a predicted value of the dependent variable:

$$\dot{y} = \dot{a} + \dot{b}\,x$$

The prediction interval puts a boundary around $\dot{y}$;

$$\dot{y} \pm A_{\varepsilon/2}$$

There is a certain level of confidence $(1 - \varepsilon)$ that the cost of a set weighing x will be in that interval. Values for $\varepsilon/2$ rather than ε are used since $\dot{y}$ is to be bounded on both sides. The value of ε can be divided by two since

under the assumptions, the probability distribution about $\dot{y}$ is normal and therefore is symmetrical. In statistical terminology, a two-tailed t distribution for constructing the intervals is used. In the case of simple regression a 100 $(1 - \varepsilon)$ - percent prediction interval for an estimated value of the dependent variable can be constructed as follows: [Ref. 1: p. 51]

$$\dot{y} \pm A_{\varepsilon/2}$$

where

$$A = (S E)\, t_{\varepsilon/2} \sqrt{\frac{n+1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

and where S E = the standard error of the estimating equation from which $\dot{y}$ was obtained,

$t_{\varepsilon/2}$ = The value obtained from a table of t-values for the $\varepsilon/2$ significance level,

n = the size of the sample,

x = the specified value of the explanatory variable used as a basis for obtaining $\dot{y}$,

$\bar{x}$ = the mean of the x's in the sample,

$\sum (x_i - \bar{x})^2$ = the sum of the squared deviations of the sample x's from their sample mean.

This prediction interval procedure can be repeated for many values of x and results plotted to obtain a 90-percent prediction interval band around the regression line, as shown in

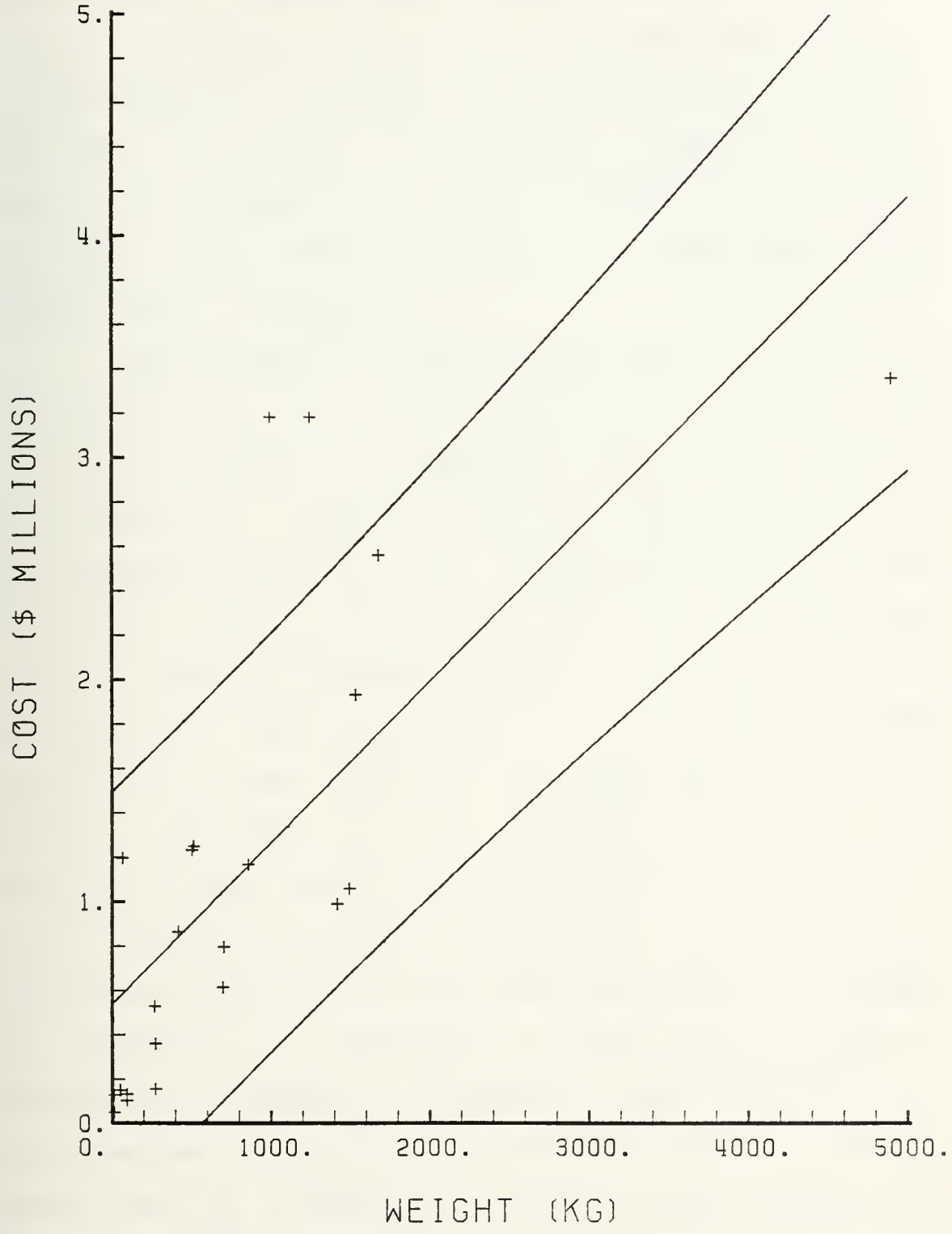


Figure 3.5 The 90-percent Prediction Interval Band for Estimated Based on Sample Data

Figure. 3.5. In this, the 90-percent confidence region is fairly wide because of the relatively large standard error of this equation. The formula for the prediction interval is such that the weight of the interval is sensitive to the size of the standard error ; large standard errors indicate that much of the cost varitation in the observed data is unexplained by the equation.

The prediction interval becomes wider as values of x farther from the mean of the sample are selected. This change in the size of the prediction interval occurs because the formulas are derived to allow for the possibility that the estimated values of a and b differ from the true values of a and b. Such a situation can occur when the sample data contain chance fluctuations that prevent the data from reflecting the true relationship that exists in the total population or when there are not sufficient data in the sample. The width of the prediction interval is also sensitive to the level of confidence that is specified and to the number of degrees of freedom. This change will make a difference in the width of the prediction interval. However, the difference in prediction interval size because of difference in degrees of freedom is more significant for small samples than for large samples; the value of t for any given level of significance becomes almost constant for degrees of freedom over 30.

B. MULTIPLE REGRESSION

To this point, simple (one explanatory variable) regression analysis has been used to examine the linear relationship between cost and weight. With the array of data shown in Table II and the logarithmic transformations of these data, multiple (more than one explanatory variable) regression analysis will now be examined. This section covers the multiple linear case. Because the sample documented in Table II contains only twenty two observations, the examination will be limited to various combinations of two rather than three explanatory variables. If additional observations were included in sample, three explanatory variables might be considered under certain circumstance; however, this number of variables used with n observations would detract from the credibility of the result. In any event, there is no great loss in limiting the number of variables to two; the essential differences between simple and multiple regression can be illustrated with the two-explanatory variable case. In the linear case, the estimating equation is of general form

$$y = a + b X1 + c X2 + d X3$$

The results for each of the possible combinations of two from the set of four explanatory variables are as follows:

$$y = 0.5353 + 0.7289 W$$

$$y = 0.6878 + 0.8209 W - 0.0943 S$$

$$y = 0.5098 + 0.5234 W + 0.0008 R$$

$$y = 0.3550 + 0.1853 S + 0.0014 R$$

$$y = 0.4826 + 0.5011 W + 0.0164 S + 0.0008 R$$

where

y = cost in millions of dollars

W = total weight in g (launch + pay load)

S = speed in mach

R = range in miles

To understand the use of t-ratios in multiple regression equations, the meaning of the multiple regression coefficients must be understood. In each case, the multiple regression coefficient shows the net effect of an explanatory variable. For example, the above equation can be interpreted as follows: For a given speed, range, a 1-kg increase in total weight will cause a $ 500 increase in cost.

As the degree of interdependence between explanatory variable increases, regression results become less stable and more indeterminant. As a consequence, the t-ratio should not be the sole test for assessing the amount of interdependence present. Further, it is not possible to give a precise cutoff point at which explanatory variables must always be considered too interdependent. A correlation coefficient of 0.9 or more between explanatory variables will almost certainly cause problems; one of 0.3 or less usually will not. [Ref. 1: p. 68]. The array of corelations among the explanatory variables should always be examined in the stages of analysis, and to the extent possible, the use of interdependent explanatory variables should be avoided.

The question arises, for cost-estimating purpose, is the multiple regression with speed and range preferable to the simple regression with weight as the explanatory variable? To find an answer, the other measures by which the regression equations are judged must be compared: the standard error of regression, the coefficient of variation, and the coefficient of determination. These are shown in Table III for each of the multiple regression for comparison with the results obtained from the simple regression. The primary concern in this comparision is between the multiple regression with speed and range and the simple regression with weight, since the speed and range equation is the only one in which both explanatory variables are significant.

TABLE III

Comparison of Multiple-linear with Simple-linear Regression Results

	Explanatory variables				
	W	W & S	W & R	S & R	W & S & R
SE	0.7547	0.7505	0.6660	0.7525	0.6338
CV	0.6563	0.6605	0.5861	0.6622	0.5578
r	0.513	0.548	0.644	0.546	0.645
DF	21	19	19	19	18

The equation above, in which weight and speed are used, appears to give slightly better results in a comparision with the other measures. However, the coefficient of the speed variable is not significant at the 10-percent level. As a consequence, the improvement is not a statistically significant one. The generalized test to determine whether the incremental improvement associated with the addition of a variable is significant uses an F-statistic. The test performed with this statistic is similar to the t-test. In this case, the null hypothesis is that the increment is not significant. The statistic used to test this null hypothesis is

$$F = \frac{\text{Increment of explained variance/ degree of freedom}}{\text{Remaining unexplained variance/ degree of freedom}}$$

This can be rewritten as

$$F = \frac{(R^2 - r^2) / 1}{(1 - R^2) / 19}$$

where R^2 = the coefficient of determination of the equation that include total weight, speed.

r^2 = the coefficient of determination of the equation with total weight alone.

Substituting the appropriate coefficients of determination in the formula for the F-statistic, we obtain

$$F = \frac{(0.548 - 0.513)}{(1 - 0.548) / 19} = 1.4712$$

This value falls short of the critical value of F, which equals 3.01 at the 10-percent level of significance. Thus, the null hypothesis is accepted [Ref. 5: p. 282]. And we conclude that the net increment in explained variance associated with the addition of speed to the equation containing weight is insufficient to establish that the improvement is not due to chance.

IV. THE LEARNING CURVE

The learning process is a phenomenon that prevails in many industries; its existence has been verified by empirical data and controlled tests. Although there are several hypotheses on the exact manner in which the learning or cost reduction can occur, the basis of learning-curve theory is that each time the total quantity of items produced doubles, the cost per item is reduced to a constant percentage of its previous cost. For example, if the cost of producing the 200th unit of an item is 80 percent of the cost producing the 100th item, and if the cost of the 400th unit is 80 percent of the cost of the 200th, and so forth, the production process is said to follow an 80-percent unit learning curve. If the average cost of producing all 200th units is 80 percent of the average cost of producing the first 100th units, the process follows an 80-percent cumulative average learning curve. There are many factors which contribute to the learning curve. These are all interrelated and, in general, no one factor can be said to be dominant over the others. The principle factors are as follows:

1. Worker efficiency
2. Method and processes
3. Total production quantity

4. Type of product

5. Lot buys

6. Tooling concepts, test equipment

The above list of relevent factors is not complete, and it tends to understate the importance of the item sometimes considered the most important--labor learning.

A. THE RELATIONSHIP BETWEEN COST AND QUANTITY

The relationship between cost and quantity may be represented by a weight equation of the form

$$y = a X^{b},$$

where X equals the cumulative production quantity. The relationship corresponds to a unit or a cumulative average learning curve according to whether y is the cost of the Xth unit or the average cost of the first X units. The constant a is the cost of the first unit produced. The exponent b, which measures the slope of the learning curve, bears a simple relationship to the constant percentage to which cost is reduced as the quantity is doubled. If S represents the decimal fraction to which cost decreases when quantity doubles, the equation becomes

$$b = \frac{\log S}{\log 2}.$$

1. Log - linear Unit Curve

If a production process follows a unit learning curve of the form $y = a x$, the cumulative cost T of producing the first n units is

$$T = a \sum_{x=1}^{n} x^{b}.$$

The cumulative average cost y of producing the first n units is then,

$$y = \frac{T}{n} = \frac{a}{n} \sum_{x=1}^{n} x^{b}.$$

2. Log - linear Cumulative Average Curve

When a production process follows a log-linear cumulative average curve rather than a unit curve, the basic functional form is still $y = a x$ but can be written $y_c = a x$, where y_c is the average cost of the first x units. The cumulative cost for producing x units is simply $y_c x$, or $a x^{b+1}$, and the unit cost is obtained from the function

$$a \{ x^{b+1} - (x - 1)^{b+1} \}.$$

and $$T = a X^{b+1}.$$

B. APPLICATIONS

The learning curve is used for a variety of purposes and in a variety of contexts; how the curve is drawn will depend on the purpose and the context. In long-range planning studies, for example, the curve must be constructed on the basis of generalized historical data, and the possible error is considerable. Empirical evidence does not support the concept of a single slope for all solid propellant missiles, all fighter aircrafts, or all spacecraft. Therefore, the practice of assuming that manufacturing hours on the airframe will follow an 80-percent curve (as was common for many years) or that electronic equipment will follow, say a 90-percent curve, can lead to very large estimating errors. For estimating to be effective, therefore, the learning curve must be established on the basis of historical data relevent to the specific problems. Such curves are equally applicable to missiles, electronic equipment, aircraft, ships and other types of equipment, but the slopes may be different for each of these.

With a small sample of data, where a learning curve is fitted to a few points, the correlation may be perfect, i.e., all the points may lie on the fitted line, but the results can still be unreliable. The points used in fitting must be sufficiently numerous and reasonably homogeneous with the points implied by extending the curve to offer a reasonable probability of sucess in predicting costs.

Whatever the basic technique, it is important to remember that on logarithmic grids the points at the right are usually more important than those that at the left. In visually fitting a line, the analyst should avoid the tendency to be unduly influenced by plot points for early lots. Early units are often incomplete because they are used for test purposes. It is equally possible that early units will include certain nonrecurring problems incident to startup and for this reason may be above the level suggested by latter plot points.

C. EXAMPLE OF LEARNING CURVE FROM BUDGET DATA

Often the only data available on a regular public basis is the budget data. This is data for total cost by year and quantity. Although this is not labor cost alone it can be used for estimating purposes. It must be adjusted to similar quantities by the learning curve. The data in this thesis came from U.S. Missile Data Book [Ref. 6] and U.S. Weapon Systems Costs [Ref. 7]. From these sources and table I for price adjustment, the data for Table II were obtained. As an example, the data for RIM-67B are shown in Table IV and the learning curve is plotted in Figure 4.1 and the calculations are shown in Appendex A. All of data for missiles in Table II were processed in a similiar way and the estimate of the cumulative average cost of 1000th unit were obtained.

TABLE IV

Plot Point of RIM - 67B

YEAR	QTY	CUM QTY	CUM AVE COST
1976	22	22	4.4421
1977	36	58	3.0591
1978	40	98	2.6326
1979	40	138	2.4300
1980	35	193	2.0713
1981	265	458	1.2077
1982	375	843	0.9544
1983	375	1218	0.9158
1984	450	1668	0.8648

Source: U.S. Missile Data Book. 1982

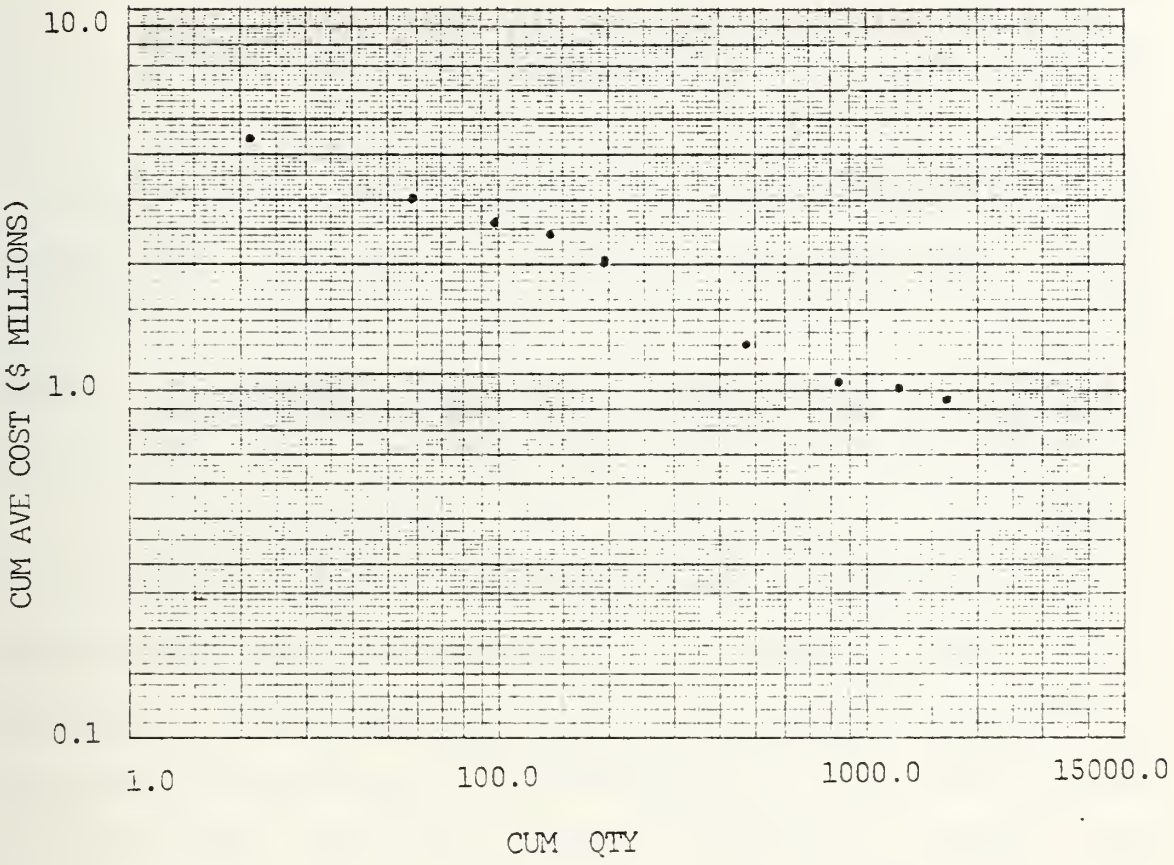


Figure 4.1 The Learning Curve of RIM-67B

V. CONCLUSION

This thesis has outlined some of the better-known methods that are used to develop cost equations. These estimating techniques range in difficulty from the simple (Simple regression) to the more difficult (Multiple regression). The latter may require the combined talents of a statistican, engineer and accountant. Statistical techniques are generally justified when the estimates are to be used in recurring decisions. The expense involved in gathering and analyzing the data for multiple regression is not usually justified if the estimate of the cost equation is to be used for only a single decision. However the missile budgeted data are available so that CERs for this area are practical.

There are some difficulties in using the statistical method for this type of study. First, there is the basic problem of obtaining a sufficient number of observations to support the distribution assumptions and to reduce the standard error. The variance of the error term is usually not known and must be estimated by the standard error. The confidence or prediction intervals depend on this measures and will be quite wide if the standard error of the estimating equation is large. The error might be reduced as the number of observations increases. Similarly, the

confidence intervals are dependent on the range of the observed values of the independent variable x. They will be relatively wide again if the range is limited and new point is outside the range of independent variable. The number of observations can sometimes be increased by using additional time periods for which cost observations are available.

I also would like to raise some questions about the validity of using the least squares criterion as a basis for cost estimation. The "least squares" estimate minimizes the sum of the deviations of actual cost observations from their estimates. By its very construction it imputes a disproportionate weight to the influence of larger deviations compared to smaller ones. This leads to the so-called "outlier" problem such as BGM-109C and MIM-104. In collecting cost observations to be included in the calculation of the parameter values, there is a tendency to discard those observations that seem to lie outside a normal trend line in order to remove a possible bias in the estimating equation. That is, their inclusion will cause the estimated cost line to tilt upward or downward in order to reduce the squared deviations between these observations and their estimates. The assumption is that outliers are merely unusual occurences and therefore should not be used to derive estimates of the normal relationship between cost item and some explanatory variable. However, so-called outliers may reflect something more basic. For example,

observations that depart from the normal trend line at the extreme ends of the range of actvity used in the analysis may reflect a nonlinear cost relationship between the cost item and the explanatory variable.

With these difficulties in statistical method, in this thesis I have introduced the estimating of the cost of U.S. missiles. It is worthwhile to further research the methodology of cost estimating for military hardware.

APPENDIX A

EXAMPLE OF CALCULATIONS FOR THE LEARNING CURVE

FY	CUM QTY	TC	CUM AVE COST
1976	22	53.7	(53.700/ 22)/0.5495 = 4.4421
1977	58	46.8	(104.205/ 58)/0.5873 = 3.0591
1978	98	50.7	(162.408/ 98)/0.6295 = 2.6326
1979	138	53.2	(230.712/138)/0.6880 = 2.4300
1980	193	50.8	(314.965/193)/0.7879 = 2.0713
1981	468	142.8	(487.372/468)/0.8623 = 1.2077
1982	843	223.4	(750.638/843)/0.9320 = 0.9544
1983	1218	310.0	(1115.435/1218)/1.000 = 0.9158
1984	1668	347.3	(1442.418/1668)/1.000 = 0.8648

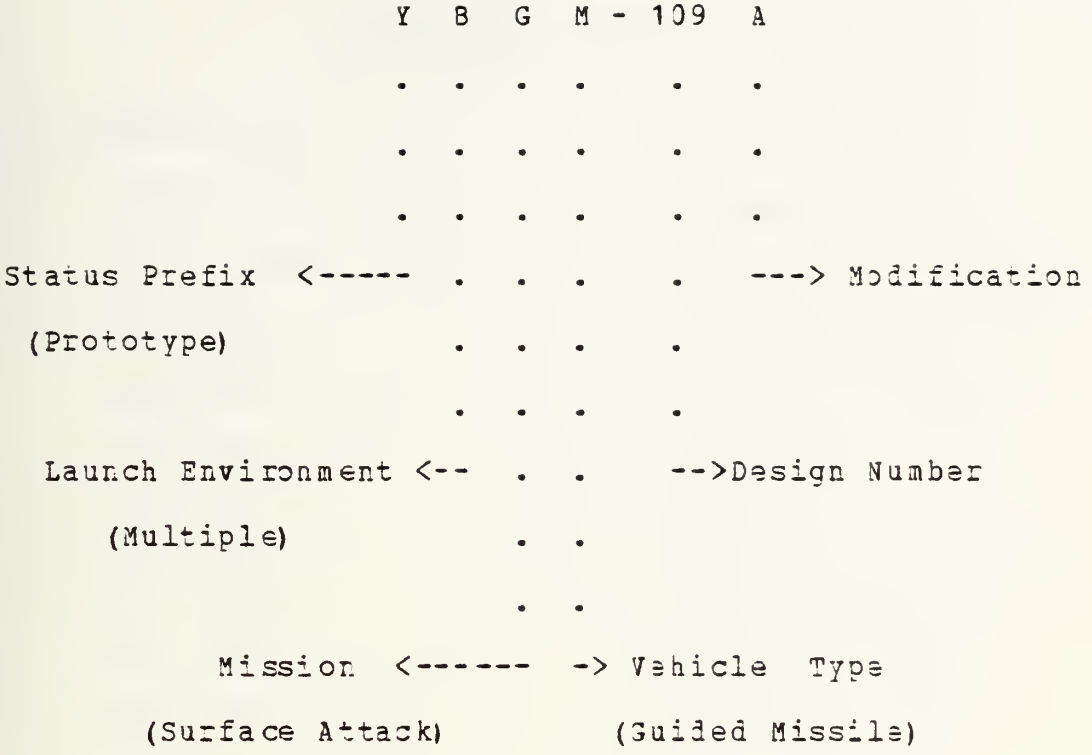
APPENDIX B

MISSILE CLASSIFICATION

Missiles are classified by the general characteristic grouping or designators. Appendix B is a cross reference listing by designator.

These grouping or designators may show in what manner a missile is used, but they will not identify a particular missile. This general classification makes use of three items: launch environment, target environment (or mission), and type of vehicle.

The first letter is used to designate the launch environment, which may be air, ground, underground, or underwater. Thus the letters are "A" for air, "G" for ground, "L" for underground or silo launched, and "U" for underwater. The second letter is used to designate the target environment or mission. This letter may be "I" for interceptor, "G" for surface target, or "Q" for drone. The third letter designates the type vehicle as "M" for missile, or "R" for rocket. An example of this general classification is illustrated:


Y B G M - 109 A
Status Prefix
(Prototype)
Launch Environment
(Multiple)
Mission
(Surface Attack)
Modification
Design Number
Vehicle Type
(Guided Missile)

APPENDIX C

DESIGNATOR LISTING

DESIGNATORS	MISSILES NAME
AIM - 7F/M	SPARROW III
AIM - 9L/M	SIDEWINDER
MIM -23B	IMPROVED HAWK
MGM -52C	LANCE
AIM -54A	PHOENIX
AIM -54C	PHOENIX
AGM -65A	MAVERICK (EO)
AGM -65D	MAVERICK (IIR)
RIM -66C	STANDARD II MR
RIM -67B	STANDARD II ER
MIM -72C	CHAPARRAL
FGM -77A	DRAGON
RGM -84A	HARPOON
AGM -86B	ALCM
AGM -88A	HARM
FIM -92A	STINGER
MIM -104	PATRIOT
BGM -109C	GLCM
BGM -109A/B	TOMAHAWK
AGM -114A	HELLFIRE
MIM -115	ROLAND II
-	PERSHING II

APPENDIX D

LAUNCH ENVIRONMENT SYMBOLS

First letter	Title	Description
A	Air	Launched from aircraft while in fight.
B	Multiple	Capable of being launched from more than one environmemt.
C	Coffin	Horizontally stored in a protective enclosure and launched from the ground.
F	Individual	Carried by one man
H	Silo Stored	Vertically stored below ground level and launched from the ground.
L	Silo Launched	Vertically stored and launched from below ground level.
M	Mobile	Launched from a ground vehicle or moveable platform.
P	Soft Pad	Partially or nonprotected in storeage and launched from the ground.
R	Ship	Launched from a surface vessel such as a ship, barge, etc.
U	Underwater	Launched from a submarine or other underwater device.

APPENDIX E

MISSILE SYMBOLS

Second Letter	Title	Description
C	Decoy	Vehicles designed or modified to confuse, deceive, or divert enemy defenses by simulating an attack vehicle.
E	Special Electronic Installation	Vehicles designed or modified with electronic equipment for communications, countermeasures, electronic raditation sounding, or other electronic recording or relay missions.
G	Surface Attack	Vehicles designed to destroy enemy land or sea targets.
I	Intercept-Aerial	Vehicles designed to intercept aerial targets in defensive or offensive roles.
Q	Drone	Vehicles designed for target, reconnaissance, or surveillance purposes.
T	Training	Vehicles designed or permanently modified for training purposes.

U	Underwater attack	Vehicles designed to destory enemy submarine or other underwater tagets.
W	Weather	Vehicles designed to observe, record, or relay data pertaining to meteorological phenomena.

APPENDIX F

VEHICLE TYPE SYMBOLS

Third Letter	Title	Description
M	Guided Missile	As the third letter in a missile designator, it identifies an unmanned, self propelled vehicle. Such a vehicle is designed to move in a trajectory which may be entirely or partially above the earth's surface. While in motion this vehicle can be controlled remotely, by homing systems, or by inertial and/or programmed guidence from within. The term "guided missile" does not include space vehicles, space boosters, or naval torpedoes, but it does not include target and reconnaissance drones.
N	Probe	The letter "N" is used to indicate nonorbital instrumented vehicles which are not involved in space missions. These vehicles are used to penetrate the space environment and transit or report back information.

R	Rocket	This identifies a self- propelled vehicle without installed or remote controll guidance mechanisms. Once launched, the trajectory or flight path of such a vehicle cannot be changed.

LIST OF REFERENCES

1. RAND, An Introduction to Equipment Cost Estimating. Prepared for: Office of the Assistant Secretary of Defense (System Analysis), 1969.

2. Booz-Allen, Applied Research Inc., Cost Management Method, 1971.

3. Bruce, M. M. and Sovereign, M.G, Parametric Cost Estimating with Application to Sonar Technology, 1973.

4. RAND, An Estimating Relationship for Fighter/ Intercepter Avionic System Production Cost, Prepared for: Office of the Assistant Secretary of Defence (System Analysis), 1966.

5. Wonnacott, T. H. and Wonnacott, R. J, Introductory Statistics, 1981.

6. Nicholas, Ted. G, U.S. Missile Data Book, Data Search Associates, Fountain Valley, CA, 1981.

7. Data Search Associates, U.S. Weapon Systems Costs, Fountain Valley, CA, 1982.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Technical Information Center Cameron Station Alexanderia, Virginia 22314	2
2.	Library, Code 0142 Naval Postgraduate School Monterey, California 93940	2
3.	Department Chairman, Code 54 Dept. of Administrative Sciences Naval Postgraduateschool Monterey, California 93940	1
4.	Professor M. G. Sovereign, Code 74 Department of Operations Research Naval Postgraduate School Montery, California 93940	3
5.	LCDR, William E. Coons, Code 55 Cj Department of Operations Research Naval Postgraduate School Monterey, California 93940	2
6.	Choo, Kyung Ho 567, Bonridong Hwawon-myun Talsung-koon Kyungbook, Korea 630-15	5
7.	Operations Research Group Army Headquarters Seoul, Korea 700-11	2
8.	System Analysis Group Ministry of Defense Seoul, Korea 131-00	2
9.	Office of the Defense Attache Embassay of The Republic of Korea 2320 Massachusetts Avenue, Northwest Washington, D.C. 20008	1
10.	Library Officer Korea Military Academy Seoul, Korea 713-55	2

Thesis
C448873 Choo
c.1

200635

The methodology of
cost estimation for
U.S. missiles.

NOV 15 85
3 DEC 86

30596

Thesis
C448873 Choo
c.1

200635

The methodology of
cost estimation for
U.S. missiles.

The methodology of cost estimation for U
3 2768 002 10383 0
DUDLEY KNOX LIBRARY